Multimodality in a Speech Aid System

Laszlo Czap, Judit Maria Pinter

Department of Automation and Communication Technology
University of Miskolc, Miskolc, Hungary

Abstract— Face to face human interaction is performed on several channels: all organs of sense are involved in conversation. Speech – as the most comfortable and natural way of human communication – is a multimodal phenomenon. Above the articulation, mimicry, gesture, posture, tactile sense and others influence perception of keynote and emotions of the speaker. Nowadays, sophisticated computers can act in a multimodal manner and gather information from different input media. Our talking head technology is being utilized in a project that aims at developing a speech assistant application, training deaf people to speak.

Keywords- talking head; face animation; expressing emotions; speech aid

I. Introduction

Intelligibility of speech can be improved by showing articulation of the speaker. This visual support is essential in a noisy environment and for people with hearing impairment. An artificial talking head can be a natural supplement to acoustic speech synthesis. A sophisticated 3D model not only can imitate the articulation, but is able to control the atmosphere of conversation. Expressing emotions is an essential element of human communication. Acceptance and attractive force of a 3D head model can be enhanced by adding emotional content to synthesized speech.

The pioneering work of face animation for modeling articulation started decades ago with 2D models [1, 2]. The development of 3D body modeling, the evolution of computers and advances in the analysis of human articulation has enabled the development of realistic models.

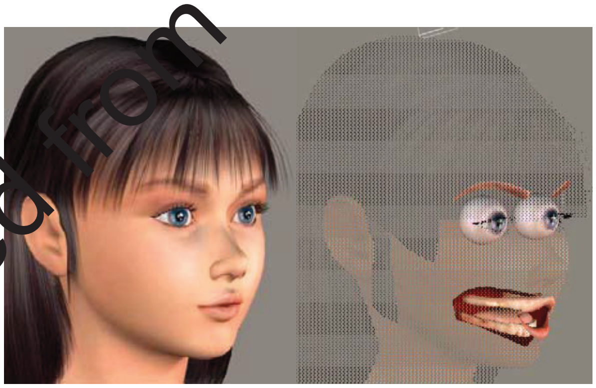


Figure 1. Photorealistic and transparent visualization

Teaching hearing impaired people to speak can be aided by an accurately articulating virtual speaker, which can make its face transparent and can show details of the movements associated with utterance (Fig. 1.).

Since the last decade, the area of visual speech has been developing dynamically, with more and more applications being developed. Existing systems are focusing on high quality modeling of articulation, but have been restricted by number of polygons issues, e.g. simulation of hair falling needs more computation effort than the speech production itself.

A visual feature database for speech-reading and for the development of 3D modeling has been elaborated and a Hungarian talking head has already been created [3]. In this research, the general approach was to use both static and dynamic analysis of natural speech to guide facial animation. A three-level dominance model has been introduced that takes co-articulation into consideration.

Each articulatory feature has been arranged in one of three classes: dominant, flexible or uncertain. Analysis of the standard deviation and the trajectory of features guide the evaluation process. The acoustic features of speech and the articulation are then linked to each other by a synchronizing process.

The aim of this research is to use the talking head technology to show the correct articulation of voices, voice transitions, words and sentences, according to the age and skillfulness of the trainee. Especially, rendering with transparent face can show the tongue movement better than a real speaker. Expressing emotions can improve the feedback of the system, praising the student or encouraging him for further exercising.

II. Talking Head

A talking head for this purpose should model photorealistic appearance and sophisticated human-like articulation close to natural movements of the speaker, triggering four additional research results:

- Pre-articulation. Prior to an utterance, a silent period is inserted – imitating breathing by opening the mouth – then the first dominant viseme is moved from the neutral starting position.
- Realizing the temporal asynchrony effect. A filtering and smoothing algorithm has been developed for adaptation to the tempo of either the synthesized or natural speech.
- Head motion, gaze, eyebrow rising, and eye blink. An algorithm has been developed to semi-randomly and manually control the former movements.
- Emotion movements. Following the definitions of Ekman, a scalable and blended algorithm was developed to express emotions.

A. Method and Material

In this line of research, a 3D transformation of a geometric surface model has been used [4, 5]. The deformation based articulation has been translated into a parametric model to overcome the restrictions of the morphing technique. Facial movements have not been carried out by deformations of the face. Instead, a collection of polygons has been manipulated using a set of parameters. This process allows control of a wide range of motions using a set of parameters associated with different articulation functions. These features can be directly matched to particular movements of the lip, tongue, chin, eyes, eyelids, eyebrows and the whole face.

The visual representation of the speech sound (mostly representing the phoneme) is called viseme. A set of visemes has fewer elements than that of phonemes. Static positions of the speech organ for the production of Hungarian phonemes can be found in seminal works [6, 7].

The features of Hungarian visemes have been constructed using the word models of [7]. The prime features of visemes have been adopted from the published sound maps and albums [6]. These features have been transformed using the properties of the articulation model by [10]. Features controlling the lips and tongue are crucial. Basic lip parameters include the opening and width, their movement are related to lip rounding. The lip opening and the visibility of teeth are dependent upon jaw movement. The tongue is specified by its horizontal and vertical position, its bending and the shape of the tongue tip (Fig. 2.).

According to the static features, the articulation parameters characteristic to the stationary section of the viseme can be set.

B. Modelling dynamic operation

The dynamic properties of conversational Hungarian have not been described yet. The current research begins to give this description. The usefulness of motion phases represented in voice albums is limited, and can be related only to the particular word given in the album.

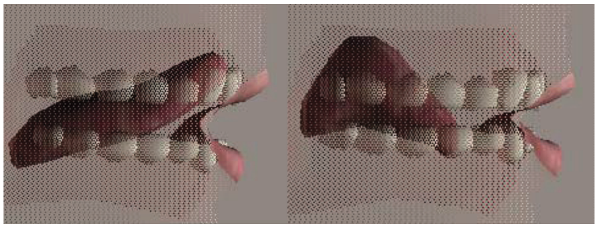


Figure 2. Illustration of tongue positions for sounds n (left) and k-g (right).

Dynamic interpretation is taken from the authors' own studies in speech reading [3]. Specifically, this work has provided the trajectories for the width and height of oral cavity and the visibility of teeth and tongue. These data provide the basis for movement between visemes.

Some parameters take on their characteristic value, while others do not reach their target value during the pronunciation. All properties of the visemes (e.g., lip shape, tongue position) have been grouped according to their dominance, with every articulation feature being assigned to a dominance class. This is different from the general approach where visemes are classified by their dominance only. For instance, the Hungarian visemes [ɟ, c, j, ɲ] are dominant on lip opening and tongue position and are dubious with respect to lip width. This grouping is based on the ranges produced by the speech reading data. Features of the parametric model can be divided into three grades:

- dominant – coarticulation has (almost) no effect on them,
- flexible – the neighboring visemes affect them,
- uncertain – the neighborhood defines the feature.

In addition to the range, the distribution of transitional and stationary periods of visible properties helps to determine the grade of dominance.

The trajectory of viseme features may also need for determining dominance classes. Fig. 3. above presents the trajectory of inner lip width (horizontal axis) and lip opening (vertical axis) of the viseme [e:]. These curves cannot be traced one by one but they go through a dense area regardless of the starting and final statuses. The dominant nature of the vowels' lip shape is manifest.

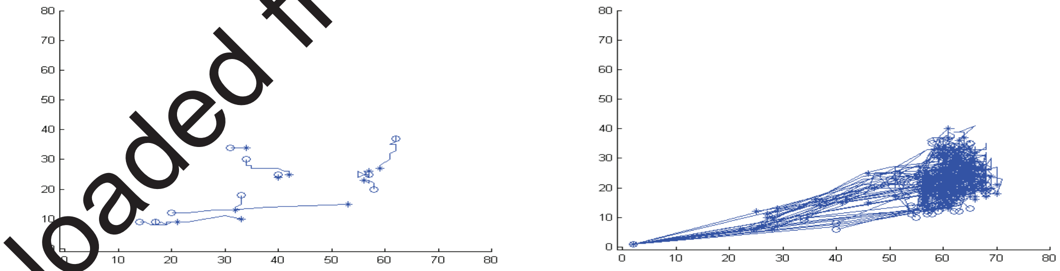


Figure 3. Trajectory of lip sizes (width and height) in pixels of viseme *e:* (above) and *h* (below).

In contrast, unsteady features do not provide a consistent pattern. The trajectory of h is shown in Fig. 3. below. (To be able to track them, only a few curves are presented.)

The dominance grade of the previously arranged parameters, which considers the deformability and context dependence of the viseme, controls the interpolation of features. Forward and backward co-articulation formulas are applied in a way that articulation parameters are influenced by less elastic properties.

C. Pre-Articulation

Prior to an utterance, there is an approximately 300 ms silence period inserted – created to imitate inhalation through the mouth – then the first dominant viseme is moved from the neutral starting status. Because of this pre-articulation movement, the sound is made as in natural speech. If the last sound of the sentence is bilabial, then the mouth would be slightly opened after the sound fade (post-articulation).

D. Audiovisual Asymmetry

Audiovisual speech recognition results indicate that performance declines monotonically and asymmetrically when there is a delay between the acoustic and visual signals.

Fig. 4. shows the audiovisual recognition rates of diphones, measured as a function of the delay between the acoustic and visual signals. The time step of audiovisual delay is 20 ms from -400 ms to 400 ms. Fig. 5. illustrates the same results in an other interpretation showing the audiovisual asymmetry obviously.

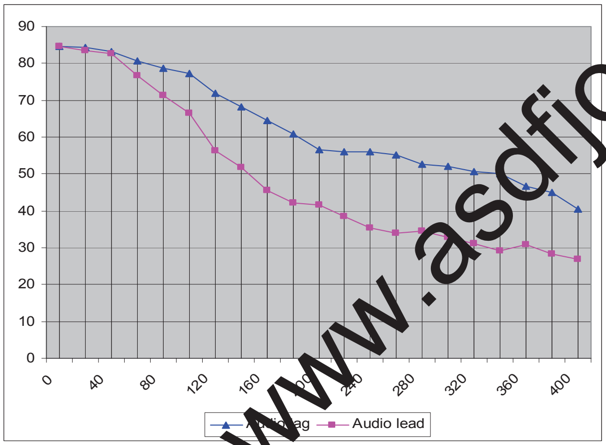
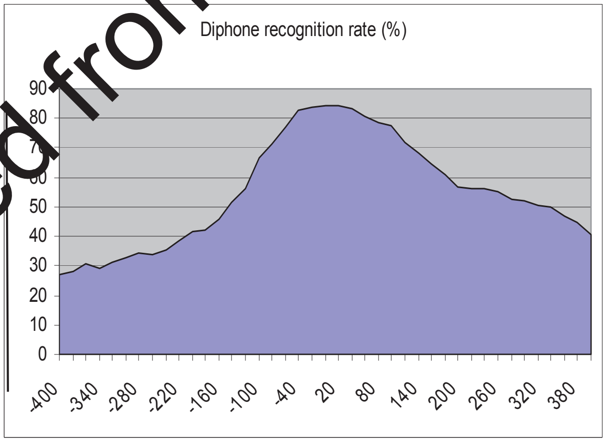


Figure 4. Declining of recognition rates. Function of audio lead and lag (ms).



Audio Delay (ms)

Figure 5. Diphone recognition rate. (Function of audio delay. Negative values mean audio leading.)

Based on the audiovisual asymmetry observed in the diphone recognition study significant improvements have been achieved in our audiovisual speech synthesis system. The filter utilizing the temporal asynchrony forwards the articulation to the acoustic signal as shown on Fig, 6.

E. Adapting to the Tempo of Speech and Filtering

During the synchronization of human or synthesized speech, we have observed various speech tempos. When speech is slow, viseme properties approach their nominal value, while fast speech is articulated with less precision in natural speech. For flexible parameters, the round off is stronger in fast speech. A median filter is applied for interpolation of flexible parameters: the values of neighboring frames are sorted and the median is chosen. A feature is created by the following steps:

- linear interpolation among values of dominant and flexible features, neglecting uncertain ones,
- median filtering is performed when flexible features are juxtaposed,
- values are then filtered by the weighted sum of the two previous frames, the actual and the next one.

The weights of the filter are fixed, thus knowledge of speech tempo is not needed. The smoothing filter refines the motions and reduces the peaks during fast speech. Considering the two previous frames, the timing asymmetry of articulation is approximated. Ref. [9] has shown that the mouth starts to form a viseme before the corresponding phoneme is audible. Filtering takes this phenomenon into consideration. Other improvements – such as inserting a permanent phase into long vowels and synchronizing phases of a viseme to a phoneme at several points – refine the articulation. Fig. 6. describes the effect of median filtering and smoothing. In this example, the slow speech has got twice as many frames as the fast one. The horizontal axis presents the number of frames, while vertical values show the amplitude of the feature.

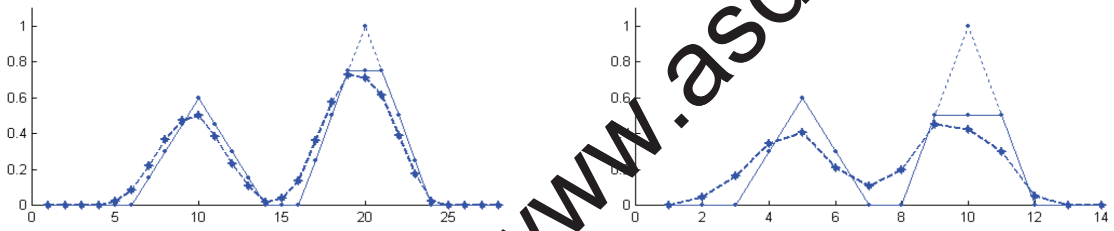


Figure 6. The interpolation of dominant (first peak) and flexible (second peak) parameters for slow (above) and fast (below) speech after linear interpolation (…), median filtering (___) and smoothing (---).

III. Improving Naturalness

Visible properties (e.g., nodding, eyebrow rising, blinking) are meaningful gestures, especially for people who are hard of hearing. Gestures can support turn taking in dialogues. For instance, the lift of eyebrows can represent paying attention, while nodding can indicate acknowledgement. An algorithm for head movement and mimicry cannot be easily produced using prosody, since the communicative context of the utterance needs to be taken into consideration. To link the automatically generated movements with the intent of the utterance, properties can be managed manually using a graphical editor (e.g., lifting eyebrows or nodding at sentence accent, or controlling the eyes to imitate a glance into a paper).

By observing the head movements of professional speakers, moderate nodding, tilting and blinking have been introduced. We have tested more than 15 minutes of speech from different announcers delivering the news on television. For each video frame, the eyes can be easily identified and traced using a conventional algorithm for obtaining motion vectors for video compression, like MPEG standards. Horizontal motion reveals panning; while vertical movement means tilting. Unbalanced movement of left and right eyes shows head inclination. A series of subjective tests have been available to fine tune these parameters.

A. Facial Gestures

Head movement, gaze, eyebrow rising, and eye blink can be regulated semi-randomly or manually. Automatic generation of facial gestures is organized in a semi-random manner. A rigid rule-based system would result in mechanical, boring and unnatural motions. Tilting and-nodding head movements are related to a short time (200 ms) average energy of acoustic signal. A downward head movement is observed for sentence accents. The bigger the average sound energy is, the higher the probability that a downward head tilting is there. Moderate and slow head turning (or pan) and side inclination are guided randomly. Amplitude of these head movements is not more than 2-3°. Gaze is controlled by monitoring head movements to keep the face looking into the camera (i.e., the observer's eyes). Vertical and horizontal head movements are created by moving both eyes in the opposite direction.

Based on observation of professional announcers and the existing literature, eyebrow movement is controlled by taking the probable strength relationship between the potential prosodic and visual cues into consideration. The interaction between acoustic intonation (F0) gestures and eyebrow movements has been observed in production, as in [10]. A preliminary hypothesis is that direct coupling is highly unnatural, but that prominence and eyebrow movement may co-occur. In our experiment, the brows have been noted to rise subtly at the beginning of declarative sentences and then it approaches the neutral position. A larger raising movement is probable to be interpreted as surprise. When emotions are expressed, inner and outer eyebrows are to be risen independently. Eyebrows can be used to disclose disgust, anger and sadness.

B. Expressing Emotions

Basic research results are published in [11, 12] and focuses on Darwin's first argument that facial expressions have evolved. When expressions are evolved, then humans must share a common, universal set. Many facial expressions are involuntary in animals, and probably in humans as well. As a consequence, much of the research has turned toward understanding what facial expressions show about the emotional state of the speaker. [13] Emotional content should be obvious to others because everyone shares the same universal assortment of expressions. In multimodal speech, we can confirm or disprove the verbal message by using gestures and body language.

Researchers have found that there are possibly six distinct emotions that are read with ease across cultures. These emotions include anger, disgust, happiness, sadness, fear, and surprise. Facial expressions signal a particular emotional state held by the speaker. This perspective leads to the conclusion that there are identifiable states that are easily perceived.

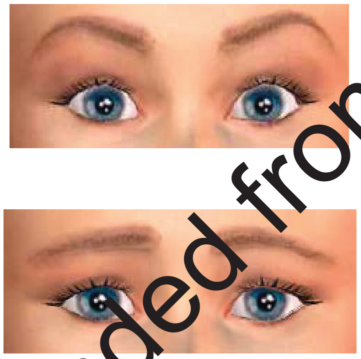


Figure 7. Eyebrows of displaying surprise and worry

A couple of features characterizing basic emotions:

Sadness: The inner corners of the eye brows are drawn up. The skin below the eye brow is triangulated, with the inner corner up. The upper eyelid inner comer is raised. The corners of the lips are down.

Disgust: The upper lip is raised. The lower lip is also raised and pushed up to the upper lip and slightly protruding. The nose is wrinkled. The cheeks are raised. Lines show below the lower lid, and the lid is pushed up but not tense. The brow is lowered, lowering the upper lid.

Happiness: Corners of lips are drawn back and up. The cheeks are raised. The lower eyelid shows wrinkles below it, and may be raised but not tense. Crow's-feet wrinkles go outward from the outer corners of the eyes.

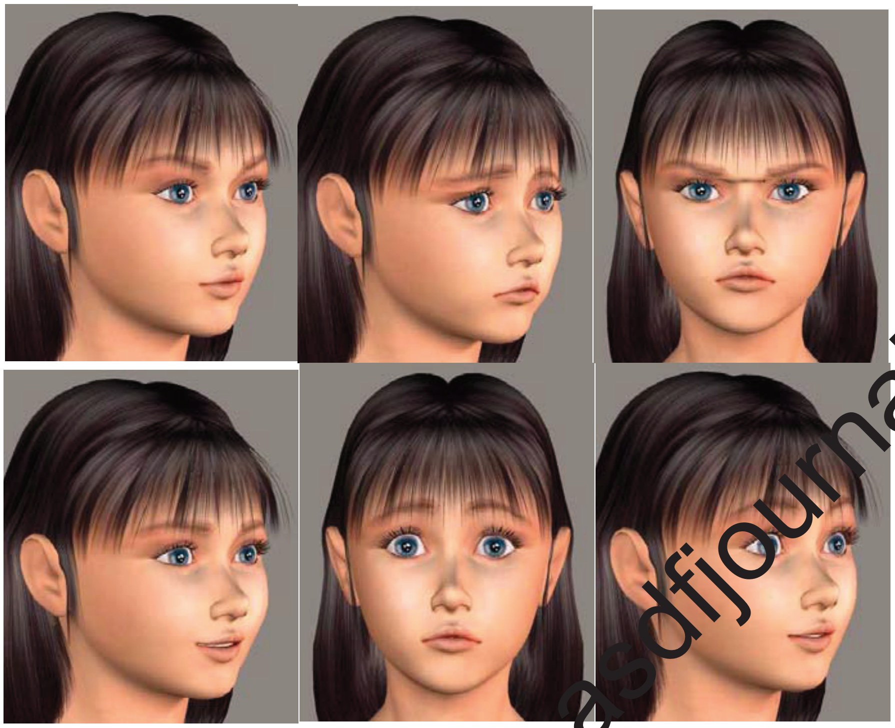


Figure 8. Expression of neutral face, sadness, disgust, happiness, fear and (nice) surprise

Surprise: The brows are raised, so that they are curved and high. The skin below the brow is stretched. Horizontal wrinkles go across the forehead. The eyelids are opened; the upper lid is raised and the lower lid drawn down; the white of the eye-the sclera-shows above the iris, and often below as well. The jaw drops open so that the lips and teeth are parted, but there is no tension or stretching of the mouth. Fig. 8. depicts six examples of emotions.

During the utterance of a sentence, facial expressions are progressing from a neutral look to the target display of emotion. Emotions can be controlled in a scalable and blended manner. E.g. 20&+30$ means 20% fear and 30% surprise, while 20*+30$ evolves 20% happiness and 30% surprise.

Acceptability is a crucial feature of a training system. It can be highly improved by a friendly appearance and kind manners of feedback when praise the student for a nice utterance, or encouraging him or her when falls behind his/her own level.

IV. Human-Machine Interactions

The first step of the system development is preparing training patterns: phonemes, transitions, words and sentences. After rendering with the talking head with different view points, zooming and lighting, we will decide the appearance of the 3D model. Testing them with trainee students and analyzing their answers, we will refine its settings. Presumably it will be different for big computer monitors and smart phones. The talking head is going to be one of the outputs of the speech assistant.

Other output is the visualized speech signal. An audiovisual transcoder converts the speech to specially designed graphic symbols. Voice can be represented real time in matrix format or the whole sample in column diagram. Both the training pattern and the actual exercise signal of the trainee student can be visualized so as to reconstruct the training pattern.

Intelligibility is highly influenced by prosody – the suprasegmental features of speech. Pitch frequency (F0) and word stress is displayed for the training sample and the result of actual exercise, to be able to compare them.

The speech assistant automatically analyzes the actual utterance and gives an assessment to the trainee student according to his/her skill level. The evaluation result has an influence on selection of the next training pattern.

Inputs of the system – above the user interactions – are

- the actual voice of the trainee to be visualized and automatically assessed,
- (optionally) the 2D or 3D video of articulation of trainee to be analyzed with the assistance of a professional trainer.

Conclusions

This paper describes the audiovisual speech synthesizer that forms the base of a larger research project that aims at creating a Hungarian audio-visual speech aid, a cognitive development system training hearing impaired people to learn speech articulation. Fine tuning of the probabilities for natural animation and avoiding mechanical, rule based repetition of gestures resulted from detailed study of speech production. Pre- and post-articulation, median filtering for adaptation to the speech rhythm and filtering for temporal asymmetry of speech production have been introduced as directions for continued fine tuning of human-like pronunciation. In this phase, further refinement of co-articulation is performed. Improving the naturalness and expressing emotions make the performance of talking head more attractive.

Sample videos can be found:
http://mazsola.iit.uni-miskolc.hu/~czap

Acknowledgment

This research was carried out as part of the TAMOP-4.2.2.C-11/1/KONV-2012-0002 project with support by the European Union, co-financed by the European Social Fund.

References

[1] E. Cosatto and H. P. Grafat, 2D photo-realistic talking head, Computer Animation, Philadelphia, Pennsylvania, 1998, pp. 103-110.

[2] M. M.Cohen and D. W. Massaro, Modeling coarticulation in synthetic visual speech, In N. M. Thalmann & D. Thalmann (Eds.) Models and Techniques in Computer Animation, Tokyo: Springer-Verlag, 1993.

[3] L. Czap, Audio-visual speech recognition and synthesis. PhD Thesis, Budapest University of Technology and Economics, 2004.

[4] L.E. Bernstein, E.T. Auer, *Word recognition in speechreading'. Speechreading by Humans and Machines*. Springer-Verlag, Berlin Heidelberg, Germany, 1996, pp. 17-26.

[5] D.W. Massaro, *Perceiving talking faces*. The MIT Press Cambridge, Massachusetts London, England, 1998, pp. 359-390.

[6] K. Bolla, *A Phonetic conspectus of hungarian*. Tankönyvkiadó., Budapest, 1995.

[7] J. Molnár *The map of hungarian sounds*. Tankönyvkiadó, Budapest, 1986.

[8] J. Mátyás *Visual speech synthesis*. MSc Thesis, University of Miskolc, 2003.

[9] G. Feldhoffer, T. Bárdi, Gy. Takács, A. Tihanyi, Temporal *asymmetry in relations of acoustic and visual features of speech*. Proc. 15th Europian Signal Processing Conf. Poznan, Poland, 2007.

[10] C. Cavé, I. Guaïtella, R. Bertrand, S. Santi, F. Harlay, R. Espesser, *About the relationship between eyebrow movements and F0 variations*. In Bunnell, H.T. and W. Idsardi (eds.), Proceedings ICSLP 96, 2175-2178, Philadelphia, PA, USA, 1996.

[11] P. Ekman, W.V. Friesen, Unmasking the Face: *A Guide to Recognizing Emotions From Facial Expressions*. MALOR BOOKS, 2003.

[12] P. Ekman, *Facial expressions* In: C. Blakemore & S. Jennett (Eds) Oxford Companion to the Body. London: Oxford University Press, 2001.

[13] N. Mana, F. Pianesi, Modelling of emotional facial expressions during speech in synthetic talking heads using a hybrid approach. Auditory-Visual Speech Processing, 2007.